論文99-36S-9-11

차량 항법용 음성인식 시스템의 구현

(Implementation of a Speech Recognition System for a Car Navigation System)

李 台 韓 *, 梁 太 榮 *, 朴 相 澤 **, 李 忠 容 *, 尹 大 熙 *, 車 日 煥 *

(Tae-Han Lee, Tae-Young Yang, Sang Taick Park, Chungyong Lee, Dae Hee Youn, and Il-Whan Cha)

요 약

본 논문에서는 차량 항법용 음성 인식을 위한 화자 독립 단독음 인식 시스템을 범용 DSP를 사용하여 구현하였으며, 잡음 처리 기술로 SNR 정규화와 RAS를 결합한 방법을 제안하여 인식 시스템의 성능을 개선시켰다. 인식 알고리즘으로서 반연속 HMM을 사용하였으며, TMS320C31을 이용하여 구현하였다. 실험에서 사용된 인식 단어는 차량 항법 시스템을 위한 명령어 69단어이며, 구현된 인식 시스템은 자동차 환경에서 녹음된 음성 데이터에 의한 인식 결과와 하드웨어 구현에 따르는 제약 조건을 동시에 고려하여 구현되었다. 주행 중에 녹음된 데이터에 대한 컴퓨터 시뮬레이션 상에서 특징 벡터 중 MFCC-CMS를 이용하고, 잡음 처리 방법으로 SNR 정규화와 스펙트럼 차감법을 결합하여 실험한 경우 최고 93.62%의 인식 성능을 보였으며, 89.93%의 인식률을 갖는 기존 방법보다 3.69%의 인식 성능 향상을 가져왔다. 제안된 잡음 처리 방법은 자동차 안에서의 SNR이 5dB이하에서 좋은 인식 성능을 보이는 것으로 나타났다.

Abstract

In this paper, a speaker-independent isolated word recognition system for a car navigation system is implemented using a general digital signal processor. This paper presents a method combining SNR normalization with RAS as a noise processing method. The semi-continuous hidden markov model is adopted and TMS320C31 is used in implementing the real-time system. Recognition word set is composed of 69 command words for a car navigation system. Experimental results showed that the recognition performance has a maximum of 93.62% in case of a combination of SNR normalization and spectral subtraction, and the performance improvement rate of the system is 3.69%. Presented noise processing method showed good speech recognition performance in 5dB SNR in car environment.

I. 서 론

음성 신호 처리 기술은 기계와 인간이 자연스럽게 많은 양의 정보를 시간이나 장소, 신체적 결함 등의 제약 없이 효율적으로 전달할 수 있는 방법으로서 꾸준히 연구되고 있다. 특히, 인간의 언어를 해석하여 적절한 행동을 수행할 수 있는 기계를 만드는 것을 목적으로 하는 음성 인식 기술은 정보 산업 분야, 디지탈 통신 분야, 가전 분야, 멀티미디어 등에 지대한 파급 효과가 있다. 현재 PCS, 셀룰라폰, 자동차, 증권 정보 안내

* 正會員, 延世大學校 電子工學科
(Dept. of Electronic Eng., Yonsei Univ.)
** 正會員, 韓國電子通信研究院 ATM 整合팀
(ATM Interface Team, ETRI)
※ 본 연구는 1997년 만도기계중앙연구소의 지원으로 연구되었습니다.
接受日字 : 1999年2月11日, 수정완료일 : 1999年7月23日

시스템[1] 등에 음성 인식 기술들이 사용되고 있다.

최근 자동차 관련 장비 중에 GPS(Global Positioning System)를 이용한 차량 항법 시스템(car navigation system)에 음성 인식 기술을 적용하려는 연구가 활발히 진행되고 있다[2]. 차량 항법 시스템은 자동차의 운행에 필요한 정보를 제공하는 시스템으로 다양한 명령어와 지명 등 복잡한 조작과정을 필요로 한다. 그러나, 운전 중의 이러한 조작은 사고의 위험성을 크게 증가시키는 원인이 된다. 음성 인식 기술은 운전자의 눈과 손을 기계조작으로부터 자유롭게 하여 운전 중 시스템 조작으로 인한 사고 위험으로부터 운전자를 보호할 수 있다. 뿐만 아니라, 음성으로 시스템을 조작함으로써 운전자의 편의성을 극대화시킬 수 있는 장점이 있다.

차량 항법용 음성 인식 시스템이 실용화되기 위해서는 자동차 잡음 환경에 강인한 음성 인식 연구가 병행되어야 한다. 음성 인식 시스템은 주변 잡음이 없는 환경에서는 만족스런 인식 성능을 얻을 수 있지만, 실제 자동차 환경은 잡음이 많이 존재하는 환경이므로 인식 시스템을 구현하는데 많은 어려움이 있다. 이와 같이 음성 인식 시스템의 성능은 학습 환경과 시험 환경에 불일치가 있는 경우 급격히 떨어진다. 이런 문제는 음성 인식 시스템이 배경 잡음과 채널 효과가 존재하는 자동차 환경에 사용될 때 필연적으로 존재하게 되므로 배경 잡음과 채널에 거의 영향을 받지 않는 강인한 음성 인식 시스템을 요하게 된다. 강인한 음성 인식을 위한 기술은 1) 음질 향상 방법[3], 2) 강인한 특징 벡터 추출[4], 3) 모델 보상 방법[5], 4) 강인한 거리 측정 방법[6]으로 크게 나눌 수 있다.

본 논문에서는 잡음 처리 방법으로 SNR 정규화[7]와 RAS(Relative Autocorrelation Sequence)[8]를 결합한 형태의 전처리 방법을 제안하였으며, 또 다른 방법으로 스펙트럼 차감법(spectral subtraction)[3]을 사용하여 실험을 하였고, 채널의 효과를 제거하기 위해 켑스트랄 평균 차감법(cepstral mean subtraction)[9]을 이용하였다. 이런 방법들을 사용하여 자동차 환경에서 실험한 결과, 기존 방법[10] 보다 향상된 인식 성능을 얻을 수 있었다. 잡음 처리 방법과 채널 효과를 제거하여 실험한 결과와 하드웨어 구현의 제약조건을 동시에 고려하여 최적의 알고리즘을 선정하여 인식 시스템을 하드웨어로 구현하였다. 인식 알고리즘은 HMM[11]중 인식률이 높은 반연속 HMM (semi-continuous HMM)[12]을 이용하였으며, 사용된 프로세서는 TI(Texas Instruments)사의 32비트 부동 소수점 프로세서인 TMS320C31[13]이고, A/D 컨버터로는 16비트 시그마 델타 A/D 컨버터인 TLC320AD56CFN[14]을 사용하여 인식 시스템을 구현하였다.

본 논문의 구성은 다음과 같다. 2장에서는 잡음 처리 방법에 관하여 설명하며, 3장에는 인식 시스템의 구성에 대해 설명한다. 4장에서는 다양한 실험과 결과에 대해 설명하며, 5장에서는 최적 인식 시스템을 하드웨어로 구현하였으며, 6장에서는 결론 및 추후 연구 과제로 끝을 맺는다.

Ⅱ. 잡음 처리 방법

최근에 멜 스케일 필터 뱅크에서 각 주파수 대역의 동적 범위를 정규화 하는 SNR 정규화 방법이 제안되어 부가적인 잡음과 채널 왜곡이 존재하는 환경에서 높은 인식률을 보였다[7]. 이 방법은 각 주파수 대역에서 측정된 동적 범위(dynamic range)에 의해 마스킹 상수를 적응적으로 사용함으로써 각 주파수 대역에서의 SNR을 정규화 하는 것이다. 그러나, SNR 정규화 방법은 잡음이 섞인 음성의 SNR이 표적 SNR보다 낮을 때는 최소의 마스킹 값이 더해져서 잡음 부분이 마스킹이 되지 않아 낮은 인식률을 보인다.

본 논문에서는 SNR 정규화 방법에서 잡음이 많은 신호에 대해 잡음을 제거하기 위한 전처리 과정으로 RAS 방법을 사용하는 것을 제안한다.

1. SNR 정규화(SNR Normalization)

자동차 환경에서의 음성 인식은 SNR 변화와 다른 녹음 채널의 영향을 다루어야한다. 따라서, 각 밴드에 대한 SNR 정규화는 잡음의 크기에 둔감한 파라메타들에 대한 효과적인 표준(criterion)을 형성한다. 그림 1은 SNR 정규화 과정의 전처리 과정을 나타내고 있다. 잡음이 없는 신호와 잡음이 존재하는 신호간의 불일치는 주로 잡음과 음성 사이의 천이 부분에서 주로 일어난다. 이런 천이에 대한 정규화는 각 주파수 대역의 동적 범위를 정규화 함으로써 이루어진다. 이 방법은 각 프레임마다 스펙트럼을 구한 후, 스펙트럼을 멜 대역으로 나누어서 선형 스펙트럼 영역의 각 주파수 대역에서 측정된 동적 범위에 의해 마스킹 상수를 적응적으로 사용함으로써 SNR을 정규화한다. 동적 범위가 정규화된 각 대역에 로그 에너지를 구해서 DCT(Discrete Cosine Transform)을 취해 12차의 MFCC(Mel-Frequen-

cy Cepstral Coefficients)를 구한다. 이렇게 함으로써 데이터와 시험 데이터의 환경을 같게 해 줄 수 있고 높은 인식 성능을 얻을 수 있다. 그러나, SNR 정규화 방법은 표적 SNR보다 낮은 잡음 환경에서는 낮은 인식 성능을 보이며, 표적 SNR을 너무 낮게 설정하면 음성의 정보가 왜곡이 생겨 인식 성능의 저하를 가져올 수 있다. 따라서, 적절하게 표적 SNR을 설정하는 것이 중요하며, 잡음을 제거하여 SNR을 높여줌으로써 잡음 부분을 마스킹하는 방법이 요구된다.

2. RAS를 이용한 SNR 정규화

SNR 정규화는 잡음이 많이 존재하는 신호의 SNR이 표적 SNR보다 낮은 환경에서는 잡음 부분이 마스킹되지 않아 낮은 인식 성능을 얻는다. 이런 SNR 정규화의 단점을 보완하기 위하여 음성 구간에 일정하게 존재하는 잡음을 감소시키기 위해 전처리 과정에 RAS방법[8]을 사용하였다. 그림 1에 제안된 방법의 전체 과정을 나타내었다.

부가적인 잡음과 채널 왜곡에 의해 오염된 입력 신호는 다음 식과 같이 표현된다.

$$y(m, n) = x(m, n) \circledast h(n) + w(m, n) \qquad 0 \le m \le M-1, \quad 0 \le n \le N-1 \tag{1}$$

여기서, m은 프레임을 나타내고, n은 한 프레임에서의 이산 시간을 나타낸다. $x(m, n)$은 잡음이 없는 음성, $y(m, n)$은 잡음이 존재하는 음성, $h(n)$은 채널의 임펄스 응답, $w(m, n)$은 부가적인 잡음, 그리고 $\circledast$은 콘벌루션 연산을 의미한다.

만약, $x(m, n)$, $w(m, n)$, $h(n)$이 서로 상관 관계가 없다면, 잡음이 존재하는 음성의 자기 상관 관계는 다음과 같이 표현된다.

$$r_{yy}(m, k) = r_{xx}(m, k) \circledast h(k) \circledast h(-k) + r_{ww}(m, k) \qquad 0 \le m \le M-1, \quad 0 \le k \le N-1 \tag{2}$$

$$r_{yy}(m, k) = \frac{1}{N-k} \sum_{j=0}^{N-1-k} y(m, j)\, y(m, j+k) \tag{3}$$

여기서, $r_{yy}(m, k)$, $r_{xx}(m, k)$, $r_{ww}(m, k)$는 잡음이 존재하는 음성, 잡음이 없는 음성, 부가적인 잡음의 단구간 자기 상관 순열을 나타내고, k는 프레임에서의 자기 상관 순열 인덱스를 나타낸다. 부가적인 잡음이 정적이라고 가정하면, $r_{ww}(m, k)$는 모든 프레임 m에 대해서 변하지 않는다. 따라서 (2)는

$$r_{yy}(m, k) = r_{xx}(m, k) \circledast h(k) \circledast h(-k) + r_{ww}(k) \tag{4}$$

가 된다. 위 식을 프레임에 대해 미분하면 잡음의 자기 상관 관계가 소거되고 이 식을 다시 근사화를 하면 다음 식과 같이 표현할 수 있다.

$$\frac{\partial}{\partial m} r_{yy}(m, k) = \frac{1}{T_L} \sum_{t=-L}^{L} t \cdot r_{yy}(m+t, k) \qquad 0 \le m \le M-1, \quad 0 \le k \le N-1 \tag{5}$$

여기서,

$$T_L = \sum_{t=-L}^{L} t^2 \tag{6}$$

이다.

위의 과정에서 구한 자기 상관 관계의 차감에 대해 FFT(Fast Fourier Transform)를 취해서 스펙트럼을 구한 뒤 SNR 정규화 과정을 수행하여 각 프레임마다 12차의 RAS-MFCC를 구한다. RAS-MFCC를 전 프레임에 대해 평균을 구해 각 프레임마다 평균을 빼주는 CMS(Cepstral Mean Subtraction)를 첨가하여 CMS-RAS-MFCC를 얻을 수 있고, 현재 프레임을 기준으로 2개의 프레임 간의 차이를 구하면 Delta-RAS-MFCC가 구해진다. 따라서, RAS와 SNR 정규화를 결합한 방법의 특징벡터는 CMS-RAS-MFCC와 Delta-RAS-MFCC, 그리고 델타 에너지(delta energy)와 델타-델타 에너지(delta-delta energy)를 결합한 것으로 이루어진다.

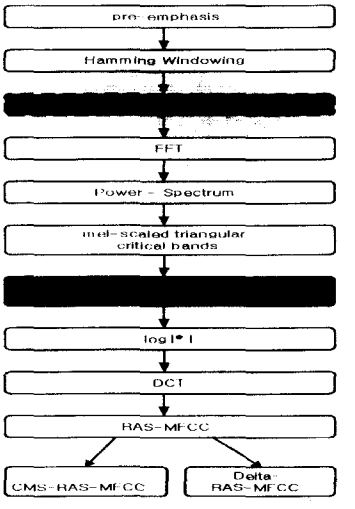


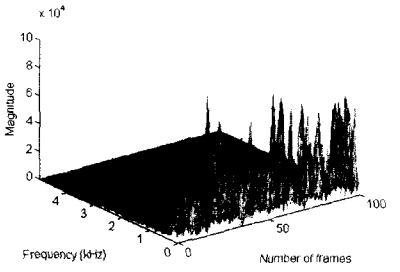
그림 1. RAS와 SNR 정규화의 결합

Fig. 1. Combination of RAS and SNR Normalization.

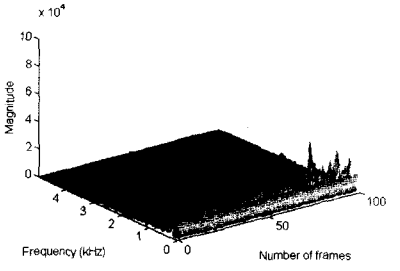
Ⅲ. 인식 시스템의 구성

1. 전처리 및 특징 벡터 추출 과정

음성 분석을 위하여 고역 통과 필터인 $1-0.95z^{-1}$을 통과시켜 프리엠프시스를 하였고, 20ms의 길이를 갖는 해밍 윈도우(Hamming window)를 사용하였다. 사용된 특징 벡터는 각 음성 프레임마다 12차의 MFCC와 현재 프레임을 중심으로 두 프레임 간격으로 델타 MFCC, 그리고 로그에너지를 구하여 현재 프레임을 기준으로 두 프레임 간격으로 델타 에너지, 델타-델타 에너지로 구성된다. 전처리 과정에서 RAS을 이용한 방법에서는 특징 벡터는 RAS-MFCC가 된다. 그리고, 본 논문에서는 채널 영향을 제거하기 위해 켑스트럴 평균 차감법(CMS)을 사용하였다.



(a)



(b)

그림 2. 100km/h의 속도로 주행중인 자동차 잡음의 스펙트로그램
(a) 잡음 신호의 스펙트로그램
(b) 고역통과 필터링된 잡음 신호의 스펙트로그램

Fig. 2. Spectrogram of the noise signal sampled in the car runing at 100km/h.
(a) Spectrogram of the noise signal
(b) Spectrogram of highpass-filtered noise signal

2. 자동차 잡음의 특징 분석

자동차 환경에서 발생하는 잡음의 형태는 그림 2와 같다. 그림 2(a)는 100 km/h로 주행중인 자동차의 햇빛 가리개에 마이크를 설치하고 1초 동안 녹음한 자동차 잡음의 스펙트로그램(spectrogram)으로 대부분의 에너지가 400Hz이하의 저주파 영역에 집중된 유색 잡음(colored noise)인 것을 알 수 있다. 본 논문에서는 이러한 잡음을 제거하기 위하여 230Hz의 고역 통과 필터를 사용하였다. 그림 2(b)는 그림 2(a) 신호를 고역 통과 필터를 통과시켰을 때의 신호의 스펙트로그램이다. 이렇게 하여 자동차 잡음이 첨가된 음성 신호에 거의 영향을 미치지 않으면서 상당 부분의 잡음을 제거할 수 있다.

3. 잡음 환경에서의 음성 구간 검출

음성 구간 검출은 입력 신호로부터 묵음과 음성을 구분하는 과정이다. 음성 구간 검출을 사용하는 인식 시스템은 구해진 음성 구간에 대해 인식을 수행하므로 인식률에 미치는 영향이 크기 때문에 정확한 음성 구간 검출이 요구된다.

간단하고 일반적인 음성 구간 검출 방법으로 단구간 에너지를 사용한다[15]. 이 방법은 단구간 에너지와 문턱치를 비교해서 끝점을 찾아낸다. 먼저, N개의 데이터를 한 구간으로 하여 로그 에너지를 구한다. 이때 신호 크기는 음성마다 달라지므로 음성의 크기에 따라 적응적으로 정규화하는 것이 필요하다. 정규화된 에너지 E_n'은 다음과 같다.

$$E_n' = 10\log_{10}E_n - Q \tag{7}$$

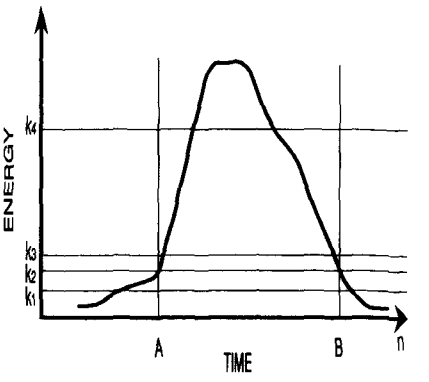


그림 3. 에너지 펄스의 시작점과 끝점을 찾는 예

Fig. 3. Example of the endpoint detection using energy pulse.

여기서, Q는 묵음 구간에서의 로그 에너지 평균 값이다.

에너지 값에 대한 문턱치 k_1, k_2, k_3, k_4(3, 5, 9, 50 dB)를 사용하여 에너지 펄스를 찾는 과정은 그림 3과 같다. 그림에서 어느 구간의 단구간 에너지가 k_1보다 크고 다시 k_1보다 작아지지 않으면서 k_2보다 크게 되는 점이 있다면 그 점을 에너지 펄스의 시작점 A로 검출한다. 또한 에너지 레벨이 다시 감소하면서 k_2보다 작고 다시 k_2보다 커지지 않으면서 k_3이하로 작아지는 점을 에너지 펄스의 끝점 B로 검출한다. 이런 과정 뒤에는 실제로 많은 에너지 펄스들이 검출되는데 펄스 지속 시간이 100ms보다 크고 최대 값이 k_4보다 큰 것을 음성 신호로 간주하고, 그 펄스의 시작점과 끝점을 찾아내게 된다.

하지만, 이러한 음성 구간 검출 방법으로는 자동차 잡음이 존재하는 상황에서의 음성 구간 검출이 매우 어려우므로 음성 크기만을 이용하는 방법으로는 정확한 음성 구간 검출이 거의 불가능하다. 따라서, 자동차 잡음 환경에서 사용될 수 있는 음성 인식 시스템 개발을 위해서는, 잡음 환경에서 사용될 수 있는 음성 구간 검출 알고리즘이 필요하다. 본 논문에서는 자동차 잡음 환경에서 정확한 음성 구간을 검출하기 위하여 음성 구간 검출 알고리즘을 크게 두 부분으로 나누어, 한 부분은 자동차 잡음을 전극 필터로 모델링한 AR(Auto Regressive) 계수를 이용하여 입력 신호를 역 필터링한 후, 신호의 크기를 이용하여 음성 구간 검출을 수행한다. 다른 부분은 자동차 잡음의 대부분이 저주파 영역에 집중되어 있으므로 230Hz의 차단 주파수를 갖는 고역 통과 필터를 통과시켜서 잡음을 제거한 신호의 크기를 이용하여 음성 구간 검출을 수행하고, 역 필터링을 이용하여 구한 음성 구간을 보정한다. 여기서, 음성 구간 검출 방식은 3개의 임계값을 사용하는 방법을 사용하였다. 역필터링을 사용하는 방법은 잡음의 스펙트럼을 모델링하여 잡음을 제거하므로 잡음을 효과적으로 제거할 수 있고 잡음의 형태에 적응할 수 있는 장점이 있다. 이 방법에 대한 블록도가 그림 4에 주어져있다.

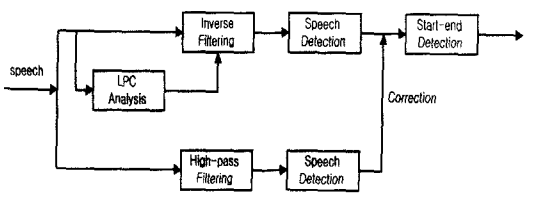


그림 4. 음성 구간 검출 방법
Fig. 4. speech-duration detection method.

4. 데이터 베이스

인식 실험에 사용된 데이터 베이스는 차량 항법 시스템에 사용되는 명령어 69단어로 구성되었으며, 인식 대상 단어는 표 1과 같다. 각 음성을 자동차에서 헤드셋(headset) 마이크로폰을 사용하여 샘플링 주파수 10kHz, 16비트로 녹음하였다.

학습 데이터는 정지한 상태에서 50명의 화자(남자 25명, 여자 25명)가 엔진을 끈 상태와 켠 상태로 각 단어를 3회씩(69단어×50명×6회=20,700단어) 발음한 음성으로 구성되었다. 시험 데이터는 19명의 화자(남자 11명, 여자 8명)가 엔진을 켠 상태에서 각 단어를 1회씩(69단어×19명×1회=1,311단어), 100 km/h로 주행중인 자동차에서 각 단어를 2회씩(69단어×19명×2회=2,622단어) 발음한 음성으로 구성되었다.

제안된 방법을 평가하기 위해 앞에서 녹음한 데이터 중 자동차 안에서 시동을 끈 상태에서 남자 13명이 각각 20단어를 3번씩 발음한 것을 사용하였다. 그 중 남자 10명의 데이터 600단어를 학습하는데 사용하였으며, 나머지 남자 3명의 데이터 180단어는 인식 성능을 평가하는 시험 데이터로 사용되었다. 자동차 안에서 100 km/h 주행 중 잡음을 샘플링 주파수 10kHz, 16비트로 녹음시켜 잡음이 없는 음성 데이터에 SNR에 따라 잡음을 첨가하여 잡음이 있는 음성 데이터를 만들었다.

5. HMM 구성

본 논문에서는 반연속 HMM을 기반으로 한 화자 독립 단독음 인식 시스템을 구현하였다. 단어 단위 인식 모델을 구성하였으며, 각 인식 대상 단어는 하나의 모델을 갖는다. 반연속 HMM에 사용되는 코드북은 위의 특징벡터 추출 과정으로부터 얻은 학습 데이터의 특징벡터들을 LBG 알고리즘[12]을 사용하여 MFCC와 델타 MFCC의 경우 128개의 코드워드, 델타 에너지와 델타-델타 에너지를 붙여 사용한 특징벡터의 경우도 128개의 코드워드를 갖도록 구성하였다. 상태 수는 10개로 고정한 경우와 1음절당 5개로 한 경우로 나누었고, 좌우 모델(left-to-right model)을 사용하였으며, 각 상태마다 4개의 혼합 확률 밀도 함수를 사용하였다.

표 1.인식 대상 단어
Table 1. Target words.

No.	단어	No.	단어	No	단어	No.	단어
0	확인	18	주간	36	계속왼쪽	54	볼륨업
1	취소	19	야간	37	계속오른쪽	55	볼륨크게
2	축소	20	반전	38	위	56	볼륨다운
3	확대	21	목적지	39	아래	57	볼륨작게
4	티비	22	출발지	40	계속위	58	크게
5	데레비	23	지도	41	계속아래	59	작게
6	텔레비전	24	지명검색	42	정지	60	밝게
7	네비게이션	25	안내정보	43	스탑	61	어둡게
8	비디오	26	검색	44	종료	62	트랙업
9	오디오	27	실행	45	끝	63	트랙다운
10	메뉴	28	현재지	46	다음화면	64	디스크
11	환경설정	29	경로편집	47	다음	65	재생
12	지도색상	30	경유지	48	페이지업	66	플레이
13	위성정보	31	귀가	49	이전화면	67	스캔
14	주행궤적	32	주행	50	이전	68	찾기
15	헤드업	33	모의주행	51	페이지다운		
16	노쓰업	34	왼쪽	52	다음채널		
17	표준	35	오른쪽	53	이전채널		

Ⅳ. 인식 실험 결과

1. 인식 결과 비교

기존 연구[10]에서는 12차의 MFCC와 12차의 델타 MFCC, 2차의 델타 에너지와 델타-델타 에너지를 합해서 3개 코드북, 가중 함수는 RPS, 512 포인트 FFT, 128개의 코드워드, 10개의 상태 수, 그리고 켑스트랄 평균 차감법을 사용해서 인식 시스템을 구현하여 89.93%의 인식률을 얻었다. 그림 5는 기존 시스템의 3번째 후보까지의 인식률을 보여준다.

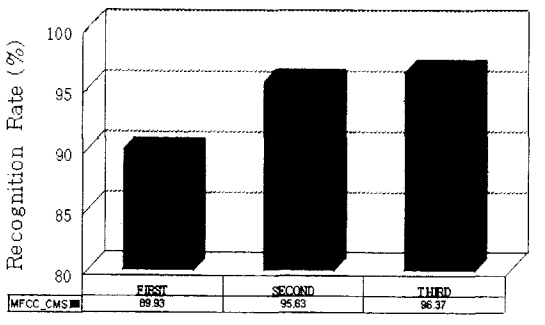


그림 5. 기존 인식 시스템의 인식률
Fig. 5. Recognition rate of the conventional system.

향상된 인식 시스템은 기존의 인식 시스템[10]을 기반으로 구성된다. 사용된 특징벡터는 12차의 MFCC와 12차의 델타 MFCC, 그리고 2차의 델타 에너지와 델타-델타 에너지를 결합한 것으로 구성되고, 가중 함수는 RPS를 사용하였다. FFT 크기는 1024개로 실험을 하였다. 코드워드는 앞 연구에서 256개와 128개가 인식률에 거의 영향을 주지 않아 메모리량을 고려하여 128개로 하였으며, 상태 수는 10개와 1음절당 5개로 구분하였다. 이렇게 Baseline을 구축을 하였고, 잡음 제거 방법으로 NSS(Non linear Spectral Subtraction)와 SNR 정규화를 사용하였으며, 채널 영향 제거 방법으로는 켑스트랄 평균 차감법을 사용하여 실험을 수행하였다.

그림 6에는 상태 수를 10개로 고정시켰을 때의 인식 결과를 나타내었다. Baseline은 91.71%의 인식률을 보였으며, 잡음 제거 방법으로 스펙트럼 차감법을 사용했을 때 92.47%, 그리고 SNR 정규화를 사용한 경우 표적 SNR이 18dB일 때가 92.52%로 가장 높은 인식 성능을 보였으나, 표적 SNR에 따라서는 인식률의 차이는 크게 나타나지 않았다. 그림에서 T18dB는 표적 SNR이 18dB임을 나타낸다. SNR 정규화에서 신호의 SNR이 표적 SNR보다 낮아 마스킹이 적절하게 되지 않는 단점을 보완하기 위해 스펙트럼 차감법을 사용하여 SNR을 높여준 후 SNR 정규화를 한 결과 인식률이 높아졌음을 알 수 있다. 특히, 표적 SNR이 12dB일 경우가 93.62%로 1.25%의 인식 성능 향상을 보여 최고의 인식 성능을 갖는 것으로 나타났다. 표적 SNR이 18dB일 경우는 가장 낮은 93.19%의 인식 성능을 보였으며, 이는 스펙트럼 차감법에 의한 깨끗한 음성의 왜곡 때문으로 판단된다.

단어 길이에 따라 상태 수를 다르게 하기 위해 1음

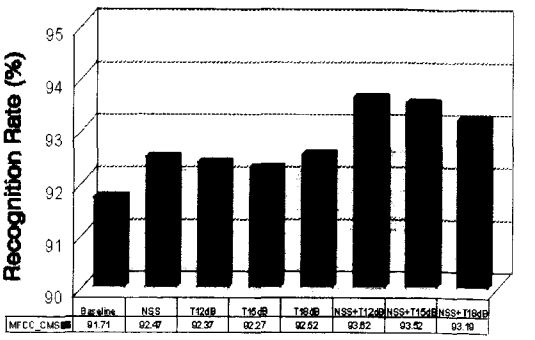


그림 6. 상태 수를 10개로 고정시킨 경우
Fig. 6. Case with 10 states.

절당 5개의 상태 수로 실험을 수행하였다. 그림 7의 결과에서 볼 수 있듯이 전체적으로 상태 수를 10개로 하는 것보다 인식 성능이 향상됨을 볼 수 있다. 표적 SNR을 18dB로 하고 스펙트럼 차감법과 SNR 정규화를 결합한 방법을 이용하였을 경우가 94.61%로 최고의 인식 성능을 보였으며 상태 수를 10개로 한 경우의 최고 93.62%보다 약 1%의 인식 성능 향상을 가져왔다.

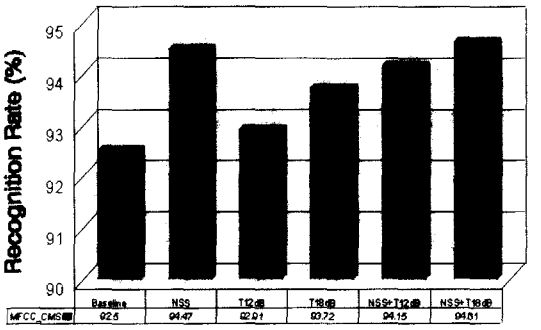


그림 7. 상태 수를 1음절당 5개로 한 경우
Fig. 7. Case with 5 states per syllable.

2. RAS를 전처리로 이용한 경우의 인식 결과

SNR이 매우 낮은 환경에서는 SNR 정규화 방법이 효과적이지 못하기 때문에 RAS 방법을 이용하여 SNR 정규화를 하는 방법을 제안한다. 잡음이 섞인 음성 신호에 자기 상관 함수를 구하여 프레임 간의 차를 구하면 일정하게 존재하는 잡음의 자기 상관 계수를 제거할 수 있다. 여기서 구한 차에 대해 Fourier 변환하여 선형 스펙트럼 영역에서 SNR 정규화를 거친 후 각 밴드에 로그값을 구해 DCT를 해서 12차의 RAS-MFCC를 구한다.

실험에서 채널 영향을 제거하기 위해 CMS를 사용하였으며, 인식 결과를 그림 8에 나타내었다. SNR 정규화 방법이 MFCC와 CMS만을 사용했을 때 보다 높은 인식 성능을 보였으나, SNR이 0dB일 때는 급격한 인식 성능의 저하를 볼 수 있다. 그러나, RAS를 사용한 SNR 정규화에서는 SNR이 0dB이고 표적 SNR이 T12dB인 경우는 SNR 정규화만 사용했을 때보다 9.4%의 인식 성능 향상을 볼 수 있다. 그림 8에서 Baseline (BL)은 MFCC와 CMS를 사용했을 경우이고, T12와 T18은 SNR 정규화 과정에서 표적 SNR이 각각 12, 18dB임을 나타내며, RAS 방법에서는 자기 상관 관계의 차감은 현재 프레임을 기준으로 4프레임 간격으로 구하였다. 잡음이 없는 음성에 대해서는 T18일 때가 최고의 인식 성능을 보이고, RAS를 사용한 SNR 정규화에서는 어느 정도 인식 성능의 저하를 볼 수 있다.

위의 인식 결과를 바탕으로 학습용 남자 25명, 여자 25명과 시험용 남자 11명, 여자 8명의 데이터를 갖고 RAS와 SNR 정규화를 결합한 방법으로 실험한 결과 표적 SNR이 12dB일 때 89.55%, 표적 SNR이 18dB일 때 88.97%의 인식률로 baseline보다 저하되는 것을 볼 수 있었다. 이는 위의 실험에서 알 수 있듯이 RAS와 SNR 정규화를 결합한 방법이 15dB이상에서는 인식률이 baseline보다 저하되는데 시험 데이타의 SNR이 평균 14.8dB 정도이기 때문이라고 생각된다.

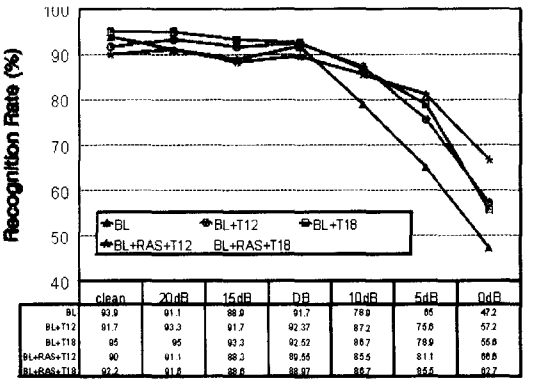


그림 8. 잡음 처리 방법과 SNR에 따른 인식 성능 비교
Fig. 8. Comparision of the recognition rate for various noise processing methods and SNR's.

V. 인식 시스템 구현

1. 구현된 시스템의 인식률

인식 실험 결과와 하드웨어 구현에 따르는 제약 조건을 동시에 고려하여 최적의 알고리즘을 선정하여 인식 시스템을 구현하였다. 표 2는 구현된 인식 시스템의 최적 알고리즘을 나타내었다. 자동차 잡음 제거 방법으로 SNR 정규화를 사용하였으며, 채널 영향을 제거하기 위해 CMS방법을 이용하였다. 음성 끝점 검출은 간단하게 로그 에너지만을 사용하였으며, 특징 벡터는 켑스트럼과 델타 켑스트럼, 그리고 델타 에너지, 델타-델타 에너지로 구성된다. 기존 연구에서 256개의 코드북과 128개의 코드북의 인식률이 크게 차이가 없어 메모리량을 고려하여 128개로 선정하였으며, FFT 크기는 기존 연구에서 1,024 포인트가 512 포인트보다 나은 인식 성능을 보여 1,024 포인트로 구현하였다.

구현된 인식 시스템의 인식 성능을 알아보기 위하여 자동차 잡음 환경에서 19명의 화자(남자 11명, 여자 8명)가 명령어 69단어를 3회씩(69단어×19명×3회=3,933단어) 발음한 데이터를 가지고 컴퓨터 상에서 시뮬레이션한 결과 93.62%의 인식 성능을 보였으며, 3번째 후보까지의 인식률은 98.04%의 인식률을 얻었다. 그림 9의 인식 결과에서 보듯이 향상된 시스템은 기존 시스템보다 3.69%의 인식 성능의 향상을 보였다.

표 2. 구현된 인식 시스템 최적 알고리즘
Table 2. Optimization of the algorithm in the implemented recognition system.

인식 단어	69 단어
인식 방법	단독음
표본화	10kHz, 16비트
프리엠퍼시스	$1-0.95z^{-1}$
윈도우	해밍 윈도우(Hamming window) 20ms size, 10ms shift
음성 구간 검출	로그 에너지
특징 벡터	12차 멜 켑스트럼 12차 델타 켑스트럼 델타 에너지, 델타-델타 에너지
잡음 제거 방법	SNR 정규화(T12dB) + SS
채널 제거 방법	켑스트랄 평균 차감법(CMS)
인식 알고리즘	반연속 HMM(SCHMM)
거리 측정 방법	RPS
코드워드 개수	128개
FFT 크기	1024 points
상태 수	10개

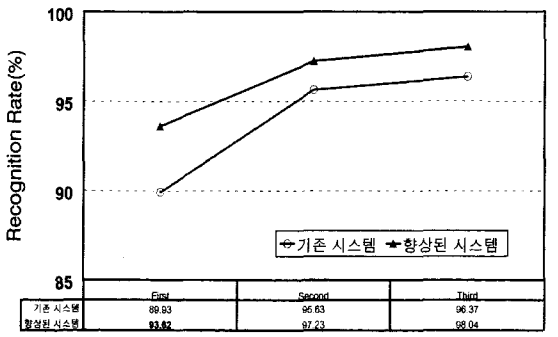


그림 9. 인식 시스템의 인식 성능 향상 비교
Fig. 9. Performance improvement of the implemented system.

2. 메모리 요구량

인식 시스템의 메모리 사용량은 코드북 크기, 특징 벡터의 차수와 개수, 단어 수, 상태 수 등에 의해 결정된다. 본 시스템에서는 69단어, 128코드워드, 12차의 특징 벡터 2개 (켑스트럼, 델타 켑스트럼), 2차의 특징 벡터 1개 (델타 에너지, 델타-델타 에너지), 그리고 상태 수를 10으로 하여 총 메모리 약 293K 워드를 사용하였다. 본 시스템을 단어 모델로 구현하였기 때문에 각 단어마다 모델 즉 초기확률, 천이확률, 관찰 확률 등을 갖고 있어 이 부분이 사용된 메모리량의 대부분을 차지하였다. 표 3에 인식 시스템의 메모리 요구량을 나타내었다.

표 3. 인식 시스템의 메모리 요구량
Table 3. Storage requirement of the implemented system.

구 분	메모리 크기 (word)
천이 확률	6,900
관찰 확률	264,960
코드북의 분산	3,328
코드북의 평균	3,328
인식에 필요한 함수	2,671
함수에 필요한 변수	17,126
스택 사이즈	256
FFT를 위한 메모리	1,536
인터럽트 벡터 테이블 0	64
인터럽드 벡터 테이블 1	62
총 메모리량	300,231(293k)

3. 소요 클럭수

인식 시스템은 성능도 우수해야 하지만 그 만큼 빠른 인식 결과를 낼 수 있어야 한다. 표 4에는 구현된 인식 시스템의 함수별 소요 클럭수를 나타내었다. 음성을 1.5초 동안 발음한다고 가정한 후 푸시버튼을 누르고 1.5초 후부터 시작하여 최종 인식 결과가 출력될 때까지 소요되는 클럭을 계산하였다.

인식 시간은 단어의 길이와 발음 속도, 그리고 끝점 검출 등에 의해 결정된다. 본 음성 인식 시스템에 사용되는 69단어들의 구성은 하나의 음절을 갖는 단어는 2개, 두 개의 음절을 갖는 단어는 28개, 세 개의 음절을 갖는 단어는 16개, 네 개의 음절을 갖는 단어는 20개, 그리고 다섯 개의 음절을 갖는 단어는 3개로 이루어진

다. 두 개의 음절을 갖는 단어와 세 개의 음절을 갖는 단어, 그리고 네 개의 음절을 갖는 단어가 대부분을 차지하기 때문에 각각 5단어를 택하여 세 번씩 보통 빠르기로 발음을 하여 평균을 내어 인식 시간을 계산하였다. 전체 평균 클럭수는 32,009,729 클럭이 소요되고, 평균 인식 시간은 1.0669초가 걸린다. 혼합 확률 밀도 함수와 로그 비터비 함수를 계산하는데 21,310,926 클럭으로 가장 많은 클럭이 소요되었으며, 다음으로는 특징 벡터 추출과 벡터 양자화가 많은 클럭이 소요되었다. 인식 시간을 감소시키기 위해서는 가장 시간을 많이 차지하는 혼합 확률 밀도 함수와 로그 비터비 함수의 계산 시간을 줄여야 한다. 혼합 확률 밀도 함수의 계산은 M(mixture)에 의해 결정됨으로 M의 갯수를 줄여 계산량을 감소시킬 수 있고, 로그 비터비 함수의 계산은 빠른 디코딩 알고리즘을 사용하여 계산량을 줄일 수 있다.

표 4. 인식 함수의 소요 클럭수
Table 4. Required clock cycles of the recognition routine.

함 수	소요 클럭수 (60MHz)	시간(sec)	점유율 (%)
끝점 검출 함수	7,170	0.000239	0.0224
SNR 정규화	695,534	0.02318	2.173
특징벡터 추출	4,785,196	0.1595	14.95
벡터 양자화	4,125,831	0.1375	12.89
확률 밀도 함수	1,085,072	0.0362	3.39
혼합 확률 밀도 함수와 로그 비터비 함수	21,310,926	0.71	66.58
합계	32,009,729	1.0669	100

VI. 결 론

본 논문에서는 음성 인식 차량 항법 시스템을 위한 화자 독립 단독음 인식 시스템을 범용 DSP 프로세서인 TI사의 TMS320C31로 구현하였으며, 잡음 처리 방법으로 SNR 정규화와 RAS를 결합한 것을 제안하였다. 구현된 시스템에서는 SNR 정규화와 NSS를 사용하여 실험한 결과 94.61%의 가장 높은 인식 성능을 얻었으며, 기존의 잡음 처리 방법을 사용하지 않고 채널 영향만을 고려한 기존 연구에서 보다 2.11%의 인식 성능 향상을 보였다. 하드웨어 구현시 메모리량을 고려하여 상태 수를 10개로 고정하였을 경우 표적 SNR이 12dB일 때 93.62%의 최고의 인식 성능으로 기존 연구 보다 3.69%의 인식 성능 향상을 보였다.

실험한 결과를 바탕으로 잡음 처리 방법을 SNR 정규화와 스펙트럼 차감법을 결합한 것을 사용하여 TMS320C31로 구현하였다. 사용된 메모리량은 약 293k 워드이며, 실제 인식 시간은 평균 1.0669초가 소요되었으며 제안된 잡음 처리 방법은 자동차 안에서의 SNR이 5dB이하에서 좋은 인식 성능을 보였다.

참 고 문 헌

[1] 최영재, 김재인, 구명완, "KT 증권정보 서비스 이용 실태 및 인식 결과 조사," 한국 음향 학회 학술발표대회 논문집, pp. 63-66, 1998년 7월.

[2] 김원구, 차일환, 윤대희, "자동차 소음 환경에서 음성 인식 시스템의 성능 향상," 음성 통신 및 신호 처리 워크샵 논문집, pp. 181-185, 1992년 8월.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 2, pp. 113-120, Apr. 1979.

[4] A. Acero and R. M. Stern, "Robust speech recognition by normalization of the acoustic space," in *Proc. IEEE Int. Conf Acoust., Speech, Signal Processing*, vol. 2, pp. 893-896, May 1991.

[5] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for MM recognition in noise," *Speech Communication*, vol. 12, pp. 231-240, 1993.

[6] B. A. Carlson and M. A. Clements, "A projection-based likelihood measure for speech recognition in noise," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 1, part 1, pp. 97-102, Jan. 1994.

[7] T. Claes and D. Van Compernolle, "SNR-normalization for robust speech recognition," in *Proc. IEEE Int. Conf Acoust., Speech, Signal Processing*, vol. 1, pp. 331-334, May 1996.

[8] K. H. Yuo and H. C. Wang, "Robust features derived from temporal trajectory filtering for speech recognition under the

corruption of additive and convolution noises," in *Proc. IEEE Int. Conf Acoust., Speech, Signal Processing*, vol. 1, pp. 577-580, May 1998.

[9] Jean-Claude Junqua and Jean_Paul Haton, *Robustness in Automatic Speech Recognition Fundamentals and Applications*, Kluwer Academic Publishers, 1996.

[10] 김지성, "차량 항법용 음성 인식 시스템 구현," 석사학위논문, 연세대학교, 1998년 8월.

[11] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-285, 1989.

[12] X. D. Huang, Y. Ariki and M. A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, 1990.

[13] Texas Instruments, *TMS320C3x User's Guide*, 1992

[14] Texas Instruments, *TLC320AD56C Data Manual*, 1996.

[15] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoint of Isolated Utterance," *Bell Syst. Tech. J.*, vol. 54, no. 2, pp. 297-315, Feb. 1975.

저 자 소 개

李 台 韓(正會員)

1997년 중앙대학교 전기공학과 학사, 1999년 연세대학교 전자공학과 석사, 1999년~현재 현대전자 이동통신 단말기본부, 주관심분야는 음성인식, CDMA

朴 相 澤(正會員)

현재 한국전자통신연구원 ATM 정합팀

尹 大 熙(正會員)

현재연세대 전자공학과 학사, Dept. of Electrical Engineering-Kansas State Univ. 석사 Dept. of Electrical Engineering-Kansas State Univ. 박사 1995.3~현재 연세 대학교 전자공학과 교수 1998.2~현재 연세대학교 신호처리연구센터 소장 주관심분야는 오디오 신호처리 (오디오 부호화) 음성 신호처리 (음성 부호화, 음성 변조, 음성 인식) 레이다 및 소나 신호처리 (적응 빔형성, 입사각추정) 적응 신호처리 (반향제거, 소음제어, 비선형 신호처리)

梁 太 榮(正會員)

1993년 연세대학교 전자공학과 석사, 1995년 연세대학교 전자공학과 석사, 1995년~현재 연세대학교 전자공학과 박사과정 재학, 주관심분야는 연속음인식, 신호처리

李 忠 容(正會員)

현재 1983.3~1987.2 연세대학교 전자공학과, 1987.3~1989.2 연세대학교 전자공학과 석사 1990.3~1991.8 연세대학교 산업기술 연구소 연구원, 1991.9~1995.12 Ph.D., Georgia Institute of Technology, Atlanta, GA 1996.2~1997.7 삼성전자 선임연구원 1997.9~현재 연세대학교 기계전자공학부 주관심 분야는 Array Signal Processing-Smart Antenna, Sonar, Position Location Speech Recognition, Communication Signal Processing

車 日 煥(正會員)

연세대 전기공학과 학사, 연세대 전기공학과 석사, 연세대학교 전자공학과 박사 1978~현재 : 연세대학교 전자공학과 교수 주관심분야는 소음진동, 건축음향, 전기음향 오디오 신호처리 (오디오 부호화) 음성 신호처리 (음성 부호화, 음성 변조, 음성 인식) 레이다 및 소나 신호처리(적응 빔형성, 입사각추정) 적응 신호처리 (반향제거, 소음제어, 비선형 신호처리)